On the Speech-To-Song Illusion: Evidence from German

Simone Falk[1], *Tamara Rathcke*[2]

[1] Department of German Language, Ludwig-Maximilians-University of Munich, Germany
[2] Department of English Language, University of Glasgow, United Kingdom
Simone.Falk@germanistik.uni-muenchen.de, T.Rathcke@englang.arts.gla.ac.uk

Abstract

The present study investigates the boundaries of speech and song from an acoustic-perceptual perspective. Using the speech-to-song illusion as a method, we tested rhythmic and tonal hypotheses to find out whether acoustic characteristics can cue the perceptual classification of a sentence by German listeners as sung or spoken. First, our results show that, despite individual differences, the speech-to-song illusion is a robust perceptual phenomenon comparable to those known in visual perception. Second, the experiment revealed that acoustic parameters – especially tonal structure – facilitate the perceptual shift from speech to song pointing to an acoustically guided decoding strategy for speech- vs. song-like signals.

Index Terms: perception, illusion, intonation, tone, rhythm, music

1. Introduction

The question whether music and language should be considered as modular entities or share common resources and processes has been hotly debated in recent discussions ([13, 14, 16]). Concerning brain modularity, the phenomenon of song is especially intriguing as it obviously combines musical and linguistic structures. The existence of musical speech or music with words seems *per se* to corroborate the hypothesis of at least some shared domains or resources ([13]). Nevertheless, the idea that singing might be exclusive to a language- or music-specific module and that during singing "music may act as a parasite of speaking" or vice versa has also been discussed ([14, 15]).

With respect to this controversy, it is worth taking a closer look at the mechanisms and acoustic premises guiding the perceptual conceptualisation of speech as opposed to song. In particular, we are interested in a perceptual phenomenon – the 'speech-to-song-illusion' (see [6, 7]) – that provides some evidence that an acoustic signal can be perceived both as speech and song. The illusion arises when a spoken utterance is presented in a loop so that the same prosodic structure with its speech-related characteristics is repeated over and over again. Surprisingly, listeners tend to perceive a shift from speech to song during the course of repetitions.

So far, the illusion has only been studied with English listeners using a single English phrase (*'(but they) sometimes behave so strangely'*) which was originally spoken in context by the author herself ([6]). Furthermore, we still do not know if the perceptual shift occurs abruptly at some point in the loop or if it is rather a continuous process. Additionally, it has been reported that the acoustics of a looped stimulus had to be unchanged since slight transposition of pitch or random permutation of syllables blocked the illusion ([7]). Adhering to a modular conception of music and speech, Deutsch et al. ([7]) proposed that the same input activated separate pathways of speech versus song processing during the task. They concluded that the perception of a phrase as speech or song could not be determined by specific music- or speech-like acoustic properties.

In contrast to Deutsch et al. ([7]), we generally assume that acoustic properties do play a role in inducing a perceptual shift from speech to song (the *Main Hypothesis*). We suggest that the stimulus used for the illusion task in [6, 7] has inherently been optimal in its acoustic layout to generate the effect. Therefore, we are going to explore whether and which acoustic features of a stimulus are better than others in inducing the shift from speech to song. More precisely, we hypothesise that two sets of signal-related features can account for (1) the fact that the auditory illusion emerges and (2) how rapidly it occurs within a loop. The *Tonal Hypothesis* assumes that the perceptual shift is predominantly induced by tonal properties of a spoken utterance, namely:

(1.a) *Target stability:* If the tonal make-up of a sentence has more stable tonal targets, the shift will occur earlier and/ or more often compared to a sentence with unstable targets ([20]).

(1.b) *Interval structure:* If the tonal properties of a sentence involve scalar relationships, the shift will occur earlier and/ or more often. Scalar structure is said to be one of the most prominent musical features that speech generally lacks ([11]).

In contrast, the *Rhythmic Hypothesis* assumes that the shift is primarily facilitated by rhythmic factors such as:

(2.a) *Accentual regularity:* If accented syllables are more regularly spaced, the shift will occur earlier and/or more frequently compared to a non-regular distribution of accents since temporal regularity of beats increases the perception of a strong rhythm (see [1]).

(2.b) *Segmental grouping:* It has been advanced by Sundberg ([19]) that intervocalic intervals constitute the smallest rhythmic units in sung speech as opposed to syllables in speech (e.g. [5]). Thus, grouping of intervocalic intervals and not of syllables will lead to an earlier shift from speech to song.

The idea of segmental grouping is exemplified in Fig. 1, showing that chunking of segments in intervocalic intervals of a sung phrase (all notes are quavers) results in a temporally more regular pattern than chunking in syllables.

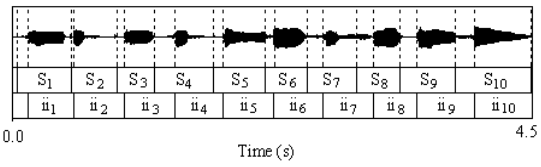

Figure 1: *Syllables (S) versus intervocalic intervals (ii) of the German phrase "Heut ist ein Fest bei den Fröschen am See" sung by an amateur female singer. Mean (S)=0.45s, SD (S)=0.1, mean (ii)=0.44s, SD (ii)=0.04.*

The present paper reports preliminary results of a series of experiments aiming to investigate the relationship between song and speech and the perceptual basis of the speech-to-song illusion.

2. Method

2.1. Materials

With respect to the *Rhythmic Hypothesis* (2.a), two sentences were chosen as experimental materials (***bold*** = accented):

- *Im Re**gal** liegen **Na**del und **Fa**den.* (English: There are needle and twine on the shelf.)
- *Im **Gar**ten blühen heute **Klee** und **Mohn**.* (English: Today, clover and poppy are blooming in the garden.)

Spoken with broad focus, accents appear regularly in the first sentence, constituting a kind of anapest (two weak units are followed by a strong beat) as opposed to the second sentence where there is no such regularity. However, both sentences have 10 syllables and 3 accents in a broad focus.

Four filler sentences were added to serve as distractors. The fillers were pairs of sentences with similar syntactic and rhythmic structures, differing from the target sentences semantically and morpho-phonologically. All sentences were read by a female speaker of Standard German and recorded in a sound-isolated booth at the Institute of Phonetics and Speech processing in Munich.

2.2. Experimental stimuli

All durational and f0-manipulations were done using *Praat* [3]. Typical f0-values of the speaker as well as her intonation patterns produced in both test utterances were used as the basis for creating tonal stimuli. To implement the tonal Target Stability Hypothesis (1.a), f0-contours between relevant f0-targets were either kept stable or changed gradually with respect to segmental landmarks. Regarding the tonal Interval Structure Hypothesis (1.b), we implemented the perfect fifth (interval of 7 st) twice in the signal: Once as an ascending interval at the first pitch-accent and again as a descending interval at the last pitch-accent of the sentence. The interval in the speech-like condition was 5.5st. A top/ base-line declination of 0.5 (st) was applied to pitch-accents/ unaccented syllables. Fig. 2 compares f0-trajectories of two stimuli, (1) music-like stimulus with scalar structure and stable f0-targets and (2) speech-like stimulus with non-scalar relationship between f0-targets and gradually changing f0-contours.

To test the rhythmic Segmental Grouping Hypothesis (2.b), the sentences had to be chunked into syllables as well as into intervocalic intervals. Due to schwa reduction in both test sentences (in /liːgən/, /blyːən/ and /faːdən/), distinguishing 10 rhythmic units for both conditions proved to be difficult. We decided to consider phrase-medial reduced forms as single rhythmic units (i.e. [liːgn] and [blyːn]) and phrase-final [faːdn̩] as two different units (see Tables 1 and 2). This decision was made with respect to the finding that phrase-final rhythmic units tend to have additional duration when compared to phrase-medial units (e.g. [12]). As shown in Table 1, intervocalic intervals all had equal duration. In contrast, in the speech-like condition (Table 2), unaccented syllables were shorter than accented ones, as it is usually the case in speech (e.g. [2]). In each sentence, duration of the final rhythmic unit accounted for final lengthening. The total length of every test sentence was 2.3 sec. Note that there were some small deviations from these target values as manipulations were done manually.

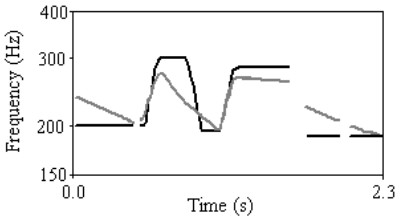

Figure 2: *F0-trajectories of the music-like tonal structure (black lines) and the speech-like tonal structure (grey lines) in the test sentence "Im Re**gal** liegen **Na**del und **Fa**den".*

The duration and tonal make-up of the filler stimuli were also manipulated in order to establish a homogenous experimental signal. Four rhythmic and/or tonal interpretations were created for each of the four fillers. The procedure described above resulted in 16 test stimuli (2 sentences x 2 durations x 2 f0-contours x 2 f0-ranges) and 16 filler stimuli (4 sentences x 4 interpretations). Each of the 32 stimuli was looped with 10 repetitions and 0.4 sec pause between them.

Table 1. *Chunking and duration of intervocalic units (in sec, song-like condition 2.b). Capitals: pitch-accented vowels.*

imr	eg	**A**ll	ignn	**A**d	el	undf	**A**d	n
img	**A**rt	nbl	ünh	eut	ekl	**E**	undm	**O**n
0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.3

Table 2. *Chunking and duration of syllabic units (in sec, speech-like condition 2.b). Capitals: pitch-accented vowels.*

im	re	g**A**l	lign	n**A**	del	und	f**A**	dn
0.20	0.20	0.35	0.20	0.35	0.20	0.20	0.35	0.25
im	g**A**r	tn	blün	heu	te	kl**E**	und	m**O**n
0.20	0.35	0.20	0.20	0.20	0.20	0.35	0.20	0.40

2.3. Procedures

The total of 32 stimuli was divided into two sets, each containing 8 test and 8 filler sentences in order to keep test duration as short as possible and to prevent subjects from hearing sentences with the same semantic content too often, i.e. no more than four times. Each test started with three practice sentences among which we presented the original English stimulus [6] and two German practice sentences. For the test phase, 16 stimuli were pseudo-randomised using one randomisation list per subject. Randomisation was guided by the following rules: (1) the test always started with a filler sentence and (2) between two test sentences with the same semantic content, at least two other (test or filler) sentences always intervened.

The subjects' task was to listen to the looped stimuli and to press a button as soon as they had the impression that the signal was no longer a spoken phrase, but had song or music-like qualities. After each loop, subjects had to confirm whether they really had perceived a change in the signal or not. Subsequently, they were asked to solve a simple mathematical equation. This was done to interrupt the test phase, to distract subjects' attention and to clear working memory load from sound imprints induced by massive repetition of the previously heard stimulus.

After the main test phase, subjects additionally performed the harmony subtest of the MBEA (*Montreal Battery of Evaluation of Amusia*, [15]). In this test, subjects had to compare musical phrases (i.e. tone sequences that differ in one pitch relationship or not) and to judge them as "same" or "different". This test was included as we expected an influence of

the subjects' musical expertise on their ability to perceive the illusion more easily. Both tests were run on computer using DMDX software ([8]). Auditory signals were presented via head-phones. One test session lasted about 25 min.

2.4. Subjects

In total, 62 native speakers of German (13 m, 49 f) aged between 19 and 46 (mean age: 24 years) participated in the experiment. 52 subjects reported to have some experience with music (singing or playing an instrument). On average, subjects had started musical training at the age of 8 and continued for about 7 years. The rest of the sample (10 subjects) had no music education. Each subject was tested separately. The sample was divided into two groups. Each group was tested once using one of the stimuli sets.

3. Results

3.1. Perception of the illusion

Overall, 60 out of 62 subjects experienced a shift from speech to song. Regarding the test stimuli, 59 out of 62 perceived the shift. The perception of the shift seems to be quite robust: in the test stimuli, 41 subjects (66 %) reported a shift in more than 50 % of the items.

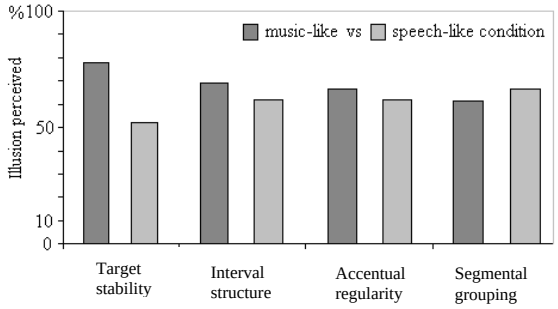

Figure 3: *Percentage of perceived speech-to-song shift depending on the four experimental factors.*

Fig. 3 gives an overview of the results showing the percentage of perceived illusions dependent on the four experimental factors. As predicted by our *Main Hypothesis*, acoustic characteristics did influence the overall perception of the illusion. In favour of the *Tonal Hypothesis*, tonal cues – in particular target stability – had the strongest effect on the perception of the illusion: stimuli with stable targets and interval structure were more likely to evoke the illusion than stimuli with unstable targets and without scalar relationships. Surprisingly, rhythmic factors did not show any obvious effects. Stimuli with regular pitch-accent distribution induced slightly more occurrences of the illusion, whereas stimuli with equal duration of intervocalic intervals were more likely to suppress rather than to facilitate the illusion which was in contrast to our prediction (see 1.a). Repeated measures ANOVA performed with four independent factors (*Target Stability, Interval Structure, Accentual Regularity, Segmental Grouping*) and dependent variable being the number of perception shifts (see Fig. 3) revealed significant results only for *Target Stability* ($F=8.5$ and $p < 0.01$) and for the interaction *Target Stability * Accentual Regularity* ($F=12.6$ and $p<0.001$, see also Fig. 5).

The histogram in Fig. 4 presents the frequency of perception shifts during the course of ten repetitions (1-10) compared to the frequency of failed illusion (0) averaged for all stimuli. Note that a similar pattern of judgement distribution was observed for each experimental condition. According to a second repeated measures ANOVA (same factors as above, dependent variable: repetition cycle), no significant differences were caused by the four tested factors, i.e. in contrast to our predictions, acoustic characteristics of the stimulus did not 'slow down' or 'speed up' the occurrence of the illusion, but rather contributed to the overall rate of perception shifts. Obviously, there was some variation in the speed of perception shift in our subject group (Median: 4th repetition). However, the illusion was mostly perceived during the third repetition of the stimulus in all conditions.

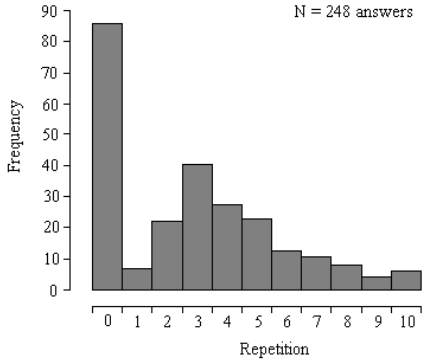

Figure 4: *Frequency of judgements for all stimuli (0=no illusion perceived, 1-10=illusion reported during the repetition).*

As discussed above, some of the test stimuli induced the illusion more frequently than others: all in all, the illusion was most often perceived in stimuli with stable targets, interval structure and syllabic grouping (with or without accentual regularity). In contrast, the illusion was least often reported for stimuli lacking these acoustic characteristics. These results are displayed in Fig. 5.

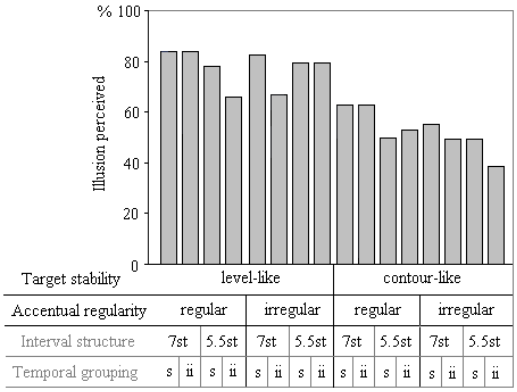

Figure 5: *Percentage of perceived speech-to-song shift for each test stimulus.*

3.2. Musicality and individual test performance

The inter-individual ranking of stimuli presented above (Fig. 5) was also intra-individually consistent. Additionally, subjects who heard the illusion more often in the overall set of test stimuli, also heard it earlier in the looped test stimuli ($n=62$**,** negative correlation between mean number of illusion occurrences per subject and mean number of repetitions, $R^2=-0.6$, $p<0.001$).

Concerning musicality, averaged MBEA-values of subjects with versus without musical education differed significantly ($p<0.01$; $t = 3.25$) showing that musicians were more sensitive to tonal changes (mean score 25 out of 30 points) than non-musicians (mean score 23 out of 30 points). However, there was no significant correlation between obtained MBEA-values and the number of illusion occurrences reported

for the test stimuli nor did MBEA-scores correlate with the tendency to perceive the speech-to-song shift earlier or later in the loop.

4. Discussion

First of all, our study confirms that the speech-to-song-illusion as described for an English sentence ([6, 7]) is robustly perceived by listeners of (at least) another intonation language, as we have demonstrated that the illusion emerges (1) in German and (2) with various sentences and manipulated speech.

In general, the results support our *Main Hypothesis*: acoustic properties of the signal did influence the perception of the illusion, and especially the overall frequency of its occurrence. So far, the study has mainly supported the *Tonal Hypothesis* since target stability significantly facilitated occurrence of the illusion. In contrast, the results were less clear in terms of the *Rhythmic Hypothesis*. We can assume that its phonetic implementation was not powerful enough to yield prominent effects in this study. However, we cannot claim that our conception of rhythm was completely wrong as we observed a significant interaction between target stability and accentual regularity. Temporal regularity of accents seems to be a secondary cue which supports the perception of the illusion. Isochronous spacing between pitch accents – i.e. an exact "beat" structure – may induce an even stronger effect and should be tested in a further experiment.

Concerning the segmental grouping condition (2.b) we encountered some problems specific to German vowel reduction processes: test sentences with schwa-elision (see 2.2) led to difficulties in chunking of intervocalic intervals. As we know, vowels in speech differ from vowels in song depending on voice register, the latter exposing more peripheral qualities in lower registers that are also used in speech ([19]). Thus, vowel reduction or undershoot in speech could have been constraining perception of the illusion. Further testing with stimuli sentences lacking schwa-vowels will shed more light on this question.

The study has shown that there were individual differences in the perception of the illusion since some subjects heard the illusion more often (and sometimes earlier) in the loop than others. In general, intra-individual performance followed the inter-individual ranking of stimuli but the study did not reveal a basis for explanation of individual performance patterns. So far, we could not find a link between the performance pattern in the illusion test and performance in the MBEA-subtest. A different battery for assessing musical abilities might be better suited to capture individual differences in the perception of the illusion.

Interestingly, when subjects did perceive a shift, they were most likely to experience it during the third repetition within the loop. This effect was also quite robust for all experimental conditions. As reported for some optical illusions (as e.g. Necker Cube, Rubin's Vase [10]) or other auditory illusions like Auditory Stream Segregation ([4]), a certain amount of time – or even a verbal instruction – is usually needed until reinterpretation of the originally perceived image is possible and a perceptual shift occurs. We assume that the repetitive structure of the signal triggers a pattern matching process in the listener that results in a strong impression of the acoustic features of a signal. As we know from other studies of the verbal domain, massive repetition of a speech signal creates a situation in which the linguistic and grammatical meaning loose their importance (the effect known as "semantic or syntactic satiation", [9, 17]). Through the establishment of a recurrent pattern, the loop becomes a rhythmically structured event in which the specific melodic and rhythmical properties of the signal may become more and more salient to the listener until they dominate the perceptual impression. In this regard, means of blocking the illusion could be e.g. variation of the tonal or rhythmic make-up of the signal (cf. [6, 7]) or the instruction to concentrate on the meaning of the sentence.

Our results do not allow direct conclusions about the modular or non-modular nature of speech and song processing. However, the fact that some acoustic properties of the signal in our experiment facilitated song perception better than others could point to a decoding strategy relying on specific acoustic cues to rate the "song-likeness" or "musicality" of a signal. This points in the direction of a modular approach, but more empirical evidence is needed to explore this issue.

5. References

[1] Auhagen, W., 2005. Rhythmus- und Tempoempfinden. In: de la Motte-Haber, H. and Rötter, G. (Eds.), *Musikpsychologie.* Laaber: Laaber, 231-250.

[2] Beckman, M., Edwards, J., 1990. Lengthening and shortening and the nature of prosodic constituency. In Kingston and M. Beckman (eds), *Papers in laboratory phonology I.* Cambridge: University Press, 152-178.

[3] Boersma, P., Weenink, D., 2001. Praat, a system for doing phonetics by computer. *Glot International 5*, 341-345.

[4] Bregman, A.S., Campbell, J., 1971. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology, 89, 244-249.*

[5] Cutler, A. 1991. Linguistic rhythm and speech segmentation. In Sundberg, J., Nord, L. and Carlson, R. (eds): *Music, Language, Speech and Brain*. Macmillan: Houndsmills &London, 157-166.

[6] Deutsch, D., 1995. *Musical Illusions and Paradoxes.* CD: Philomel Records.

[7] Deutsch, D., Lapidis, R., Henthorn, T., 2008. The speech-to-song-illusion. *J. Acoust. Soc. America, s.124,* 2471.

[8] Forster, K.I. and Forster, J., 2003. DMDX: A Windows Display Program with Millisecond Accuracy. *Behavior Research Methods, Instruments, & Computers 35 (1),* 116-124.

[9] Francom, J.C., 2009. *Experimental Syntax: exploring the effect of repeated exposure to anomalous syntactic structure – evidence from rating and reading tasks*. Dissertation, University of Arizona.

[10] Gregory, R.L., 1971. Visual illusions. In Held, R. (ed.) *Image, object and illusion.* San Francisco: Freeman, 48-58.

[11] Krumhansl, C.L., 2005. Musikalische Bezugssysteme. In Stoffer, T.H. and Oerter, R. (Eds), *Allgemeine Musikpsychologie*. Berlin, New York: de Gruyter, 267-306.

[12] Lehiste, 1973. Rhythmic units and syntactic units in production and perception. *J. Acoust. Soc. Amer. 54(5).* 1228-1234.

[13] Patel, A., 2008. *Music, language and the brain.* Oxford: University Press.

[14] Peretz, I., in press. Music, language and modularity in action. In Rebuschat, P., Rohrmeier, M., Hawkins, J. and Cross, I. [Eds], *Language and Music as cognitive systems.* Oxford: University Press.

[15] Peretz, I., Champod, S. and Hyde, K. (2003). Varieties of Musical Disorders: The Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Sciences 999,* 58-75.

[16] Peretz, I. Coltheart, M., 2003. Modularity of music processing. *Nature neuroscience, 6,* 688-691.

[17] Pilotti, M., Antrobus, J.S. and Duff, M. (1997). The effect of presemantic acoustic adaptation on semantic 'satiation'. *Memory and Cognition 25 (3),* 305-312.

[18] Sundberg, J., 1987. *The Science of the Singing Voice.* Illinois: Northern Illinois Press.

[19] Sundberg, J., 1989. Synthesis of Singing by rule. In Mathews, M. V. and Pierce, J. R. [Eds], *Current Directions in Computer Music Research.* MIT Press, 45-55.

[20] Zatorre, R.J., Belin, P. and Penhune,V.B., 2002. Structure and function of auditory cortex: music and speech. *Trends in cognitive Science 6,* 37-46.